

Multicategory reclassification statistics for assessing improvements in diagnostic accuracy

JIALIANG LI*

*Department of Statistics and Applied Probability, National University of Singapore,
Singapore 117546, Singapore
Duke-NUS Graduate Medical School, Singapore 169857, Singapore
Singapore Eye Research Institute, Singapore 168751, Singapore
stalj@nus.edu.sg*

BINYAN JIANG

*Department of Statistics and Applied Probability, National University of Singapore,
Singapore 117546, Singapore*

JASON P. FINE

Departments of Biostatistics and Statistics, University of North Carolina, Chapel Hill, NC 27517, USA

SUMMARY

In this paper, we extend the definitions of the net reclassification improvement (NRI) and the integrated discrimination improvement (IDI) in the context of multicategory classification. Both measures were proposed in Pencina *and others* (2008. Evaluating the added predictive ability of a new marker: from area under the receiver operating characteristic (ROC) curve to reclassification and beyond. *Statistics in Medicine* **27**, 157–172) as numeric characterizations of accuracy improvement for binary diagnostic tests and were shown to have certain advantage over analyses based on ROC curves or other regression approaches. Estimation and inference procedures for the multiclass NRI and IDI are provided in this paper along with necessary asymptotic distributional results. Simulations are conducted to study the finite-sample properties of the proposed estimators. Two medical examples are considered to illustrate our methodology.

Keywords: Area under the ROC curve; Integrated discrimination improvement; Multicategory classification; Multinomial logistic regression; Net reclassification improvement.

1. INTRODUCTION

Diagnostic and classification tasks are encountered in medical practice where we need to accurately differentiate the disease-present and the disease-absent status of a patient. Sometimes the classification may involve more than two categories which need to be treated separately. For example, one of the motivating examples of this paper is a classification for synovitis where patients from five distinct disease categories

*To whom correspondence should be addressed.

are examined, each requiring a different patient management strategy. Health-care procedures are decided based on the determination of the most likely status of the patient. Biomarkers are often used to predict the patient's disease status, drawing upon a set of well-established statistical tools for evaluating diagnostic accuracy.

Typically, in binary classification, one employs statistical methods based on the receiver operating characteristic (ROC) analysis and provides the estimated ROC graph and/or summary measures of the ROC graph such as the area under the curve (AUC) or the partial AUC (Zhou and others, 2002). Such methods vary the threshold used to classify individuals as diseased and non-diseased and then combines the resulting sensitivities and specificities across all possible thresholds. The extensions of these methods to multiclassification have been recently proposed. Specifically, we may use the hypervolume under the multidimensional ROC manifold (HUM) as an extension of the AUC to evaluate the classification accuracy for any biomarker in a multiclass problem (Mossman, 1999). The HUM has an interpretation akin to the AUC where a large HUM value indicates a high classification accuracy. The inference procedure for the HUM has been discussed in Nakas and Yiannoutsos (2004) for ordered polychotomous responses and Li and Fine (2008) for unordered polychotomous responses. Shiu and Gatsonis (2012) provides a review of ROC-type methods for multiclass problems.

While ROC-based measures have been widely adopted, it has been argued by many authors (Pepe and others, 2004; Pencina and others, 2008) that such measures may not be good criteria to quantify improvements in diagnostic accuracy when the added value of a new predictor to an existing model is of interest. Such analyses are critically important in the development of predictive models based on biomarkers, where the added value of markers which may be expensive to obtain must be weighed against the associated financial costs. The interpretation of the AUC provides an indirect assessment of the predictive performance of a model. Thus, the gain with a new predictor may be unclear. A related issue is that the AUC measures may be relatively insensitive to the addition of predictors in certain regions of the AUC space. To address these limitations, Pencina and others (2008) proposed two novel criteria based on reclassification in order to directly quantify the extent to which a new predictor improves classification performance: the net reclassification improvement (NRI) and the integrated discrimination improvement (IDI). These measures have met with widespread success in the medical literature, with many practitioners preferring their ease of interpretation versus ROC-based measures. For additional discussion of these recent developments, we refer the reader to Pencina and others (2011, 2012).

In this paper, we extend the reclassification indices to multiclassification problems, providing an alternative to HUM-based analyses. The multiclassification definitions of the NRI and the IDI are presented in Section 2.1, with their inferential procedures described in Section 2.2. The use of such measures in model building is discussed in Section 2.3, with an eye toward high-dimensional data structures, where the number of predictors may be much larger than the sample size. Extensive numerical studies are conducted. Simulations are reported in Section 3, with two real data examples presented in Section 4, including the synovitis example mentioned previously and a microarray example, where it is important to select a small number of the most important expression biomarkers for the prediction of cancer (Li and Fine, 2008; Ma and Song, 2011, among others).

2. METHODS

2.1 Accuracy parameters

Consider a set of predictors $\Omega = \{X_1, \dots, X_p\}$. Suppose that we have a sample of subjects with measurements of all X_j ($j = 1, \dots, p$). The goal is to use them to construct a meaningful statistical model for predicting the multiclassification outcome Y which takes values from $\mathcal{Y} = \{1, 2, \dots, M\}$. We define the binary random variable $Y_m = I(Y = m)$ and let the prevalence for the m th category be $\rho_m = E(Y_m) = P(Y = m)$.

Suppose that a model \mathcal{M}_1 is constructed based on a set of predictors $\Omega_1 \subset \Omega$. Such a model \mathcal{M}_1 can generate a probability vector $\mathbf{p}(\mathcal{M}_1) = (p_1(\mathcal{M}_1), \dots, p_M(\mathcal{M}_1))$ for each subject where $\sum_{m=1}^M p_m(\mathcal{M}_1) = 1$. Decision makers may assign a subject to one of the M categories according to the greatest component in the probability vector. One may quantify the accuracy of \mathcal{M}_1 based on Ω_1 by the following multicategory correct classification probability (CCP):

$$\text{CCP} = \sum_{m=1}^M \rho_m \text{CCP}_m, \quad (2.1)$$

where each CCP_m is the CCP for the m th-category. For the model \mathcal{M}_1 , we write

$$\text{CCP}_m(\mathcal{M}_1) = P\{p_m(\mathcal{M}_1) = \max \mathbf{p}(\mathcal{M}_1) | Y = m\}, \quad m = 1, \dots, M. \quad (2.2)$$

Now suppose that more variable(s) are included and we construct a model \mathcal{M}_2 that is based on a set of predictors $\Omega_2 \supset \Omega_1$. We use the nested-structure notations as they are widely discussed in the literature. We note that there are plenty of cases where the accuracy improvement occurs among non-nested models. Our proposed methods can apply with slight modification.

The newly constructed model \mathcal{M}_2 generates another probability vector $\mathbf{p}(\mathcal{M}_2) = (p_1(\mathcal{M}_2), \dots, p_M(\mathcal{M}_2))$ for each subject where $\sum_{m=1}^M p_m(\mathcal{M}_2) = 1$. Again, decision makers may assign the subject according to the greatest value of this probability vector, and the m th-category accuracy of \mathcal{M}_2 based on Ω_2 can be quantified by

$$\text{CCP}_m(\mathcal{M}_2) = P\{p_m(\mathcal{M}_2) = \max \mathbf{p}(\mathcal{M}_2) | Y = m\}, \quad m = 1, \dots, M. \quad (2.3)$$

The overall accuracy improvement from \mathcal{M}_1 to \mathcal{M}_2 may be summarized as

$$\mathbf{T} = \sum_{m=1}^M w_m \{\text{CCP}_m(\mathcal{M}_2) - \text{CCP}_m(\mathcal{M}_1)\}, \quad (2.4)$$

where w_m are positive weights for the m th category. When $M = 2$, the two CCPs are usually called the sensitivity and specificity of the model-based test. The \mathbf{T} measure thus quantifies the overall increase of the weighted sum of sensitivity and specificity. When equal weights are used for the two categories, \mathbf{T} is simply the difference of Youden's index between the two models. We shall call (2.4) the *reclassification improvement* (RI) since it reflects how the accuracy changes after a reclassification.

The RI measure has well-known limitations in assessing improvements in diagnostic accuracy and has not been widely adopted in practice (Pencina and others, 2008). Here, we propose to extend the NRI, which has recently been studied as an alternative to the RI for binary classification. The multicategory extension of the NRI is

$$\mathbf{S} = \sum_{m=1}^M w_m P\{p_m(\mathcal{M}_2) = \max \mathbf{p}(\mathcal{M}_2), p_m(\mathcal{M}_1) \neq \max \mathbf{p}(\mathcal{M}_1) | Y = m\}. \quad (2.5)$$

When $M = 2$ and $w_m = \frac{1}{2}$, \mathbf{S} is equivalent to the NRI given in Pencina and others (2008). We refer to \mathbf{S} as the NRI in this article since it indicates the probability that added markers in \mathcal{M}_2 lead to correct classification of subjects who are incorrectly classified using the smaller model \mathcal{M}_1 . We note that, in the two-category classification, the decision can be based on whether the class probability exceeds $\frac{1}{2}$ with equal priors on the two categories.

The IDI can be generalized to multiple categories by noticing the connection between the IDI in binary classification problems and R^2 (Cox and Wermuth, 1992; Menard, 2000; Tjor, 2009). The interpretation

and computation of R^2 , also called a coefficient of determination, has been discussed for binary logistic regression models. Simply speaking, the value of R^2 is the fraction of the total variation explained by the model. For linear regression models, R^2 is closely related to the correlation coefficient and the ANOVA F -test, while for binary regression, it is closely connected to the probabilities of correct classification.

Let $\mathbf{R}^2(\mathcal{M}_j) = (R_1^2(\mathcal{M}_j), \dots, R_M^2(\mathcal{M}_j))$ be an M -dimensional vector with

$$R_m^2(\mathcal{M}_j) = \frac{\text{var}(Y_m) - E\{\text{var}(Y_m|\mathcal{M}_j)\}}{\text{var}(Y_m)} = \frac{\text{var}\{p_m(\mathcal{M}_j)\}}{\rho_m(1 - \rho_m)}. \quad (2.6)$$

The second equality follows because $E(Y_m|\mathcal{M}_j) = p_m(\mathcal{M}_j)$. It has been shown in [Pepe and others \(2008\)](#) that the increase in R^2 for binary classification ($M = 2$) from model \mathcal{M}_1 to model \mathcal{M}_2 is equivalent to the IDI in [Pencina and others \(2008\)](#). A natural adaptation of the R^2 definition of the IDI to the multicategory set-up is

$$\mathbf{R} = \sum_{m=1}^M w_m \{R_m^2(\mathcal{M}_2) - R_m^2(\mathcal{M}_1)\}. \quad (2.7)$$

We refer to (2.7) as the IDI in this paper, similarly to the binary case. The multicategory IDI (2.7) reduces to that in [Pencina and others \(2008\)](#) when $M = 2$ and equal weights $w_1 = w_2 = \frac{1}{2}$ are used.

The choice of weights in the definitions of the NRI and the IDI may depend on the goal and design of the study. When aiming for the overall test accuracy to differentiate multiple classes, it is natural to weigh all categories equally; on the other hand, as pointed out in [Pencina and others \(2011\)](#), sometimes it is useful to reward some categories with higher weights when savings associated with correct classification of such categories outweigh other categories. When cost-efficiency information is available, we can incorporate them easily in the inference for the weighted NRI and IDI. There are also other practical considerations that invoke unequal weights and one can run a Bayesian prior elicitation procedure to construct reasonable weights ([Li and Fine, 2010](#)).

2.2 Estimation theory

Suppose we obtain a sample $\{X_{i1}, \dots, X_{ip}, Y_i : i = 1, \dots, n\}$. We denote the class sample size $n_m = \sum_{i=1}^n I(Y_i = m)$ for the m th category. We assume that $n \rightarrow \infty$ and $n_m/n \rightarrow \rho_m > 0$ for all m . One may fit the candidate models \mathcal{M}_1 and \mathcal{M}_2 from the preceding subsection using one's method of choice. The main requirement is that the method provide estimated probability assessment vectors for each model. Using the fitted models, one may then estimate the two reclassification accuracy measures RI and NRI by

$$\begin{aligned} \hat{\mathbf{T}} = & \sum_{m=1}^M \frac{w_m}{n_m} \sum_{i=1}^n \{I(\hat{p}_{mi}(\mathcal{M}_2) = \max \hat{\mathbf{p}}_i(\mathcal{M}_2), Y_i = m) \\ & - I(\hat{p}_{mi}(\mathcal{M}_1) = \max \hat{\mathbf{p}}_i(\mathcal{M}_1), Y_i = m)\} \end{aligned} \quad (2.8)$$

and

$$\hat{\mathbf{S}} = \sum_{m=1}^M \frac{w_m}{n_m} \sum_{i=1}^n I\{\hat{p}_{mi}(\mathcal{M}_2) = \max \hat{\mathbf{p}}_i(\mathcal{M}_2), \hat{p}_{mi}(\mathcal{M}_1) \neq \max \hat{\mathbf{p}}_i(\mathcal{M}_1), Y_i = m\}, \quad (2.9)$$

where $\hat{\mathbf{p}}_i(\mathcal{M}_j) = (\hat{p}_{1i}(\mathcal{M}_j), \hat{p}_{2i}(\mathcal{M}_j), \dots, \hat{p}_{Mi}(\mathcal{M}_j))$ is the estimated membership probability for the i th subject based on the j th model. In practice, they can be obtained from fitting a multinomial logistic regression model to the data and then outputting the predicted probabilities from the fitted model. We note that when the models are consistently estimated, $\hat{\mathbf{p}}_i(\mathcal{M}_j)$ is consistent to $\mathbf{p}_i(\mathcal{M}_j)$ and therefore, by the law of large numbers, $\hat{\mathbf{T}}$ and $\hat{\mathbf{S}}$ are consistent to \mathbf{T} and \mathbf{S} , respectively. Furthermore, by using the central limit

theorem, we can show that $\sqrt{n}(\hat{\mathbf{T}} - \mathbf{T}) \rightarrow_d N(0, \sigma_T^2)$, where

$$\sigma_T^2 = \sum_{m=1}^M \frac{w_m^2}{\rho_m} (a_m + b_m - 2c_m) - \sum_{i=1}^M \sum_{j=1}^M w_i w_j (a_i - b_i)(a_j - b_j), \quad (2.10)$$

with $a_m = \text{CCP}_m(\mathcal{M}_1)$, $b_m = \text{CCP}_m(\mathcal{M}_2)$, $c_m = P(p_m(\mathcal{M}_1) = \max \mathbf{p}(\mathcal{M}_1), p_m(\mathcal{M}_2) = \max \mathbf{p}(\mathcal{M}_2) | Y = m)$, and $\sqrt{n}(\hat{\mathbf{S}} - \mathbf{S}) \rightarrow_d N(0, \sigma_S^2)$, where

$$\sigma_S^2 = \sum_{m=1}^M \frac{w_m^2}{\rho_m} d_m - \sum_{i=1}^M \sum_{j=1}^M w_i w_j d_i d_j, \quad (2.11)$$

with $d_m = P(p_m(\mathcal{M}_1) \neq \max \mathbf{p}(\mathcal{M}_1), p_m(\mathcal{M}_2) = \max \mathbf{p}(\mathcal{M}_2) | Y = m)$. We note that the terms following $\sum_{i=1}^n$ in $\hat{\mathbf{S}}$ and $\hat{\mathbf{S}}$ are not independent. We obtain independent sums only after changing the order of summation and then are able to apply the central limit theorem. In practice, all the components involved in the variance expressions can be consistently estimated by plugging in the sample version of the probabilities.

The multicategory IDI can be estimated by using the following formula:

$$\hat{\mathbf{R}} = \sum_{m=1}^M \frac{w_m}{n_m(1 - n_m/n)} \sum_{i=1}^n \{[\hat{p}_{mi}(\mathcal{M}_2) - \overline{\hat{p}_m(\mathcal{M}_2)}]^2 - [\hat{p}_{mi}(\mathcal{M}_1) - \overline{\hat{p}_m(\mathcal{M}_1)}]^2\}, \quad (2.12)$$

where $\overline{\hat{p}_m(\mathcal{M}_j)} = n^{-1} \sum_{i=1}^n \hat{p}_{mi}(\mathcal{M}_j)$.

We can also show that (2.12) is consistent to \mathbf{R} for a large n by noting that, for a large sample, $\hat{p}_{mi}(\mathcal{M}_j)$ is consistent to $p_{mi}(\mathcal{M}_j)$, and the average squared distance to the mean $n^{-1} \sum_{i=1}^n \{\hat{p}_{mi}(\mathcal{M}_2) - \overline{\hat{p}_m(\mathcal{M}_2)}\}^2$ is consistent to $\text{var}\{p_m(\mathcal{M}_j)\}$. The consistency then follows from the law of large numbers. As with RI and NRI, one may further show that $\sqrt{n}(\hat{\mathbf{R}} - \mathbf{R}) \rightarrow_d N(0, \sigma_R^2)$, where

$$\begin{aligned} \sigma_R^2 = & \sum_{j=1}^M \sum_{k=1}^M \frac{w_j w_k}{\rho_j \rho_k (1 - \rho_j)(1 - \rho_k)} \{E[(p_j(\mathcal{M}_1) - \mu_j(\mathcal{M}_1))^2 - (p_j(\mathcal{M}_2) - \mu_j(\mathcal{M}_2))^2] \\ & \times ((p_k(\mathcal{M}_1) - \mu_k(\mathcal{M}_1))^2 - (p_k(\mathcal{M}_2) - \mu_k(\mathcal{M}_2))^2) - E[(p_j(\mathcal{M}_1) - \mu_j(\mathcal{M}_1))^2 - (p_j(\mathcal{M}_2) \\ & - \mu_j(\mathcal{M}_2))^2] E[(p_k(\mathcal{M}_1) - \mu_k(\mathcal{M}_1))^2 - (p_k(\mathcal{M}_2) - \mu_k(\mathcal{M}_2))^2]\}, \end{aligned} \quad (2.13)$$

and $\mu_m(\mathcal{M}_j) = E(p_m(\mathcal{M}_j))$. All the moments involved in the variance expression can be readily estimated by using empirical moment estimators. The variance can then be estimated by the plug-in method.

The above parameter estimation and variance estimation formula are implemented in the software R and the code is downloadable at <http://www.stat.nus.edu.sg/~stalj>. Although the variance formula (2.11) and (2.13) look complicated in the above presentation, our experiences with simulation and real data analysis suggest that they can be evaluated instantly following the point estimation by using our code. These formula allow inferences to be carried out much faster than a resampling-based approach. An advantage of the resampling method is that the sampling variability in the estimation of the probability vector may be formally accounted for in the inference.

2.3 Model-building procedure

For biomedical data with ever-growing dimensionality and complexity, we often face the challenge $p \gg n$ and cannot afford using all p markers for the construction of a feasible prediction model. We now propose a

procedure to select important predictors for regression analysis with multicategory response. Specifically, we adopt a forward selection algorithm by using the NRI and the IDI as the selection criteria. The model-building algorithms are similar for the two criteria. We thus choose to present the detailed procedure for the NRI only.

We start with a null model. At the first step, we fit regression models with a single covariate for all X_j , $j = 1, \dots, p$ and evaluate the NRI for each X_j . The variable that gives the highest NRI value is chosen at this stage. At the second step, we fit regression models with the previously selected predictor and another predictor in the remaining set. We evaluate the NRI again for each model and then select the best model according to the highest NRI. The selection procedure proceeds until a stopping rule is satisfied.

We consider two stopping rules in this paper. Rule I stops the model-building procedure when a pre-specified number of predictors or pre-specified proportion of all predictors is achieved. Rule II stops the model-building procedure when a pre-specified full-model accuracy is achieved, for example, 90% of the overall CCP. When resources are limited and only a fixed number of markers can be fully investigated in a study, we may consider Rule I and retain a relatively parsimonious model; on the other hand, when there is sufficient support that allows us to examine as many markers as possible, we may target a very high accuracy and choose Rule II. As demonstrated in the microarray analysis in Section 4.2, it may be possible to achieve 100% accuracy in some applications. Simulation studies included in supplementary material available at *Biostatistics* online find satisfactory performance of this forward selection method.

3. SIMULATION ANALYSIS OF ESTIMATION CONSISTENCY

In this section, we used simulation studies to examine the performance of our proposed estimators \hat{S} and \hat{R} for the NRI and the IDI, respectively. We consider six different scenarios.

Case 1. We first consider a three-category response with the following multinomial logistic structure:

$$\log \frac{p_2}{p_1} = -0.5 + 1.5X_1 + 3X_2, \quad \log \frac{p_3}{p_1} = -1 + 2X_1 + 4X_2,$$

where $p_j = P(Y = j)$ for $j = 1, 2, 3$. We generate (X_1, X_2) from a multivariate normal distribution with mean $(1, 1)$ and covariance matrix $\Sigma = (\sigma_{i,j})_{1 \leq i,j \leq 2}$. We let $\sigma_{11} = \sigma_{22} = 1$, and $\sigma_{12} = \sigma_{21} = 0$ and let $\mathcal{M}_1 = \{X_1\}$, $\mathcal{M}_2 = \{X_1, X_2\}$.

Case 2. The same as Case 1 except that we let $\sigma_{12} = \sigma_{21} = 0.2$.

Case 3. Next we increase the number of covariates and consider the following three-category response model:

$$\begin{aligned} \log \frac{p_2}{p_1} &= -2 + 0.25X_1 + 0.5X_2 + 0.75X_3 + 1X_4 + 1.25X_5, \\ \log \frac{p_3}{p_1} &= -4 + 2.5X_1 + 2.25X_2 + 2X_3 + 1.75X_4 + 1.5X_5. \end{aligned}$$

We generate (X_1, X_2, \dots, X_5) from a multivariate normal distribution with mean $(1, 1, 1, 1, 1)$ and covariance matrix $\Sigma = (\sigma_{i,j})_{1 \leq i,j \leq 5}$. We set $\Sigma = \text{diag}\{1, 1, 1, 1, 1\}$ and let $\mathcal{M}_1 = \{X_1\}$, $\mathcal{M}_2 = \{X_1, X_2, X_3, X_4, X_5\}$.

Case 4. The same as Case 3 except that we let $\sigma_{ij} = 0.1, i \neq j$. This imposes a compound symmetry dependence structure for the covariates.

Table 1. Performance of \hat{S} and \hat{R} over 1000 simulations

	S	Avg.SD _S	\hat{S}	SE _{\hat{S}}	CR _S	R	Avg.SD _R	\hat{R}	SE _{\hat{R}}	CR _R
Case 1										
$n = 150$	0.263	0.067	0.271	0.055	0.977	0.263	0.053	0.279	0.052	0.925
$n = 300$	0.263	0.047	0.269	0.040	0.973	0.263	0.036	0.272	0.036	0.937
Case 2										
$n = 150$	0.238	0.059	0.246	0.057	0.955	0.252	0.049	0.262	0.050	0.935
$n = 300$	0.238	0.042	0.246	0.041	0.952	0.252	0.034	0.258	0.034	0.951
Case 3										
$n = 150$	0.282	0.094	0.322	0.076	0.977	0.319	0.088	0.385	0.085	0.938
$n = 300$	0.282	0.062	0.298	0.051	0.977	0.319	0.054	0.346	0.052	0.946
Case 4										
$n = 150$	0.302	0.083	0.317	0.071	0.976	0.334	0.076	0.384	0.077	0.930
$n = 300$	0.302	0.056	0.305	0.048	0.969	0.334	0.049	0.356	0.049	0.944
Case 5										
$n = 150$	0.226	0.057	0.251	0.046	0.975	0.178	0.039	0.203	0.038	0.922
$n = 300$	0.226	0.039	0.237	0.032	0.975	0.178	0.025	0.186	0.024	0.942
Case 6										
$n = 150$	0.223	0.053	0.249	0.046	0.953	0.183	0.036	0.206	0.036	0.900
$n = 300$	0.223	0.037	0.238	0.033	0.957	0.183	0.024	0.194	0.024	0.939

S and R are the true value of NRI and IDI computed using Monte Carlo; avg.SD_S and avg.SD_R are mean of estimated SEs of \hat{S} and \hat{R} computed based on (2.11) and (2.13) over 1000 simulations; \hat{S} and SE _{\hat{S}} are the mean and standard deviation of the estimated NRI over 1000 simulations; \hat{R} and SE _{\hat{R}} are the mean and standard deviation of the estimated IDI over 1000 simulations; CR_S = (number of simulations the true NRI falling into the interval $[\hat{S} - 1.96SE_{\hat{S}}, \hat{S} + 1.96SE_{\hat{S}}])/1000$; CR_R = (number of times the simulations the true IDI falling into the interval $[\hat{R} - 1.96SE_{\hat{R}}, \hat{R} + 1.96SE_{\hat{R}}])/1000$.

Case 5. Next we consider a five-category response by using the following multinomial logistic structure:

$$\begin{aligned} \log \frac{p_2}{p_1} &= -4 + 3X_1 + 3.5X_2 + 4X_3 & \log \frac{p_3}{p_1} &= -3 + 2.5X_1 + 3X_2 + 3.5X_3, \\ \log \frac{p_4}{p_1} &= -2 + 2X_1 + 2.5X_2 + 3X_3 & \log \frac{p_5}{p_1} &= -1 + 1.5X_1 + 2X_2 + 2.5X_3. \end{aligned}$$

We generate (X_1, X_2, X_3) from a multivariate normal distribution with mean $(1, 1, 1)$ and covariance matrix $\Sigma = \text{diag}(1, 1, 1)$, $\mathcal{M}_1 = \{X_1\}$, $\mathcal{M}_2 = \{X_1, X_2, X_3\}$.

Case 6. The same as Case 5 except that we set $\sigma_{12} = \sigma_{21} = \sigma_{23} = \sigma_{32} = 0.1$.

For each case, we simulated 1000 data sets and carried out the estimation procedures for the NRI and the IDI introduced in the previous section. The numerical results are summarized in Table 1. Eyeballing Table 1, we notice that both \hat{S} and \hat{R} perform very well in estimating S and R, respectively, in all cases. The variance formula for the NRI and the IDI also provide very close approximation to the sampling variability of the estimators. The coverage rates are close to 95% and improve as the sample size increases. We note that in some cases the coverage of the IDI is slightly lower than the nominal level. The undercoverage issue may be caused by the fact that we do not account for the sampling variability of the regression estimation. Another factor is that normal approximation is not appropriate when the IDI is close to zero. After acknowledging the estimation uncertainty, Kerr and others (2011) derived an approximate χ^2 distribution under the null for the two-category classification. Another alternative approach is to consider the bootstrap.

4. EXAMPLES

We first consider a medical study with five categories to illustrate the applications of the NRI and the IDI for a practical assessment of accuracy improvement. A second analysis utilizes data from a genetic study to demonstrate how the NRI and the IDI can be used for variable selection in high-dimensional data analysis. Since the HUM has previously been employed for model selection in multiclass set-ups, we explicitly compare the HUM to the new measures, with a focus on interpretation and model selection issues.

4.1 Tissue biomarkers of synovitis

Synovitis is the inflammation of the synovial membrane and may occur in association with arthritis as well as lupus, gout, and other conditions. [Krenn and others \(2006\)](#) and [Slansky and others \(2010\)](#) described a scoring approach to evaluate the grading of the histological severity of synovitis. In their data set, the primary classification outcome involved five different categories: normal healthy control; Post-traumatic Arthropathy (PtA); Osteoarthritis (OA); Psoriatic Arthritis (PsA); Rheumatoid Arthritis (RA). The sample included a total of 583 synovial tissue specimens taken from 33 normal, 29 PtA, 221 OA, 42 PsA, and 341 RA subjects, respectively. Details of the sample collection procedures were included in earlier publications ([Krenn and others, 2006](#); [Slansky and others, 2010](#)).

Three biological measurements are commonly used to predict the patient disease status: the lining thickness, the inflammatory infiltrates, and the stromal density. Each of the three biomarker components was graded on a scale from zero to three in the sample. The accuracy of the three biomarkers for pairwise binary classification and five-category classification were reported in [Krenn and others \(2006\)](#) and [Slansky and others \(2010\)](#). We use our proposed approach to evaluate the relative improvement for increasing the marker numbers in a statistical model. For the ease of presentation and comparison, we denoted the lining thickness as X_1 , the stromal density as X_2 , and the inflammatory infiltrates as X_3 in this paper. We first quantified the diagnostic accuracy for each component and then combined them with the multinomial logistic regression to further improve the diagnostic accuracy.

The estimated NRI and IDI are reported in Table 2. The models with a single covariate are not nested and cannot be compared using the NRI and the IDI. Instead, we report their CCP in the NRI column and their R^2 values in the IDI column. The second marker X_2 has the highest CCP and IDI and is considered as the most accurate one. The CCP indicates that over 65% of the sample were correctly classified by X_2 , while the IDI indicates that approximately 12% of the overall variability of the five-category response may be attributed to X_1 . The other two markers have relatively inferior performance, with X_3 being the next most accurate marker and followed by X_1 . This finding is consistent to the previous observations in [Slansky and others \(2010\)](#) using ROC-based diagnostic accuracy measures, with the estimated HUMs for X_1 , X_2 , and X_3 being 0.0005, 0.0140, and 0.0075, respectively.

It is of interest to combine X_2 with other marker(s) to improve the overall diagnostic accuracy. NRI results suggest that a two-marker model combining X_2 with X_3 (NRI = 0.0385) may result in a larger improvement in classification accuracy than with X_1 (NRI = 0.0030). The NRI values indicate that the improvement on a model-based net reclassification rate due to including X_3 is more than 10 times of that due to including X_1 . We note that marker X_3 has the second highest CCP 0.5866 in the one-marker model. The standard errors (SEs) in the brackets allow us to construct the Wald-type tests. In this case, it seems that the NRI improvement for the addition of X_3 is highly significant, whereas that for adding X_1 is not.

On the other hand, IDI results show a similar preference for X_3 . The IDI improvement for including X_3 (IDI = 0.0302) is three times of that for including X_1 (IDI = 0.0102), indicating that the addition of X_3 could explain much more variability in the response. Using the SEs, we carried out the Wald tests and found that both IDI increases were significantly different from zero.

Table 2. *Estimated NRI and IDI and SEs for three synovitis biomarkers: lining thickness (X_1), stromal density (X_2), and inflammatory infiltrates (X_3)*

Model(s)	NRI (SE)	IDI (SE)
X_1	0.5746 (0.0081)	0.0555 (0.0168)
X_2	0.6518 (0.0078)	0.1174 (0.0304)
X_3	0.5866 (0.0085)	0.0751 (0.0180)
$X_1 + X_2$ versus X_2	0.0030 (0.0017)	0.0102 (0.0021)
$X_3 + X_2$ versus X_2	0.0385 (0.0056)	0.0302 (0.0024)
$X_1 + X_2 + X_3$ versus $X_2 + X_3$	0.0013 (0.0009)	0.0062 (0.0015)

For models with only one marker, NRI is the overall model CCP and IDI is the model R^2 . For others, NRI is the NRI value and IDI is the increase in IDI for the two nested models (X_3).

In practice, the final model including three markers is always considered as it yields the highest CCP and IDI. Moving from model $\{X_2 + X_3\}$ to model $\{X_1 + X_2 + X_3\}$, the NRI is not significant (Wald test $|0.0013/0.0009| < 1.96$), while the IDI is highly significant (Wald test $|0.0062/0.0015| > 1.96$). These divergent results require careful interpretation because the two measures need to be construed from different perspectives. Adding X_1 in the last stage may not significantly improve the number of correct classifications from the previous model. However, it may still contribute a significant amount of information and help explain the overall variability of the response.

The results in Table 2 are interesting to investigators as how the model accuracy evolved with the added complexity was clearly presented and the relative contribution of each marker to the overall classification was fully demonstrated. The earlier findings in [Slansky and others \(2010\)](#) reported the accuracy for individual markers (in the absence of information from other markers) but failed to inform the added accuracy of new markers in the presence of other already selected marker(s).

4.2 Leukemia classification

We next consider data from leukemia patients used in [Golub and others \(1999\)](#). The data are from a study of gene expression in three types of acute leukemias, acute lymphoblastic leukemia arising from T-cells (ALL T-cell), ALL arising from B-cells (ALL B-cell), and acute myeloid leukemia (AML). The data set contains 8 ALL T-cell samples, 19 ALL B-cell samples, and 11 AML samples. Each sample contains 7912 gene expression values obtained from Affymetrix high-density oligonucleotide microarrays. Our data set is downloaded from

<http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi>.

We consider evaluating the accuracy of the biomarkers for their ability to differentiate the three classes. The single gene that gives the highest CCP is the 1184th gene in the data file, with a CCP value 0.8421. We then use the methods in this paper to select a second gene which maximizes the accuracy improvement. By using the NRI, the best gene is the 2216th gene with an NRI value 0.2145 and SE 0.0678 (P -value 0.0015). Adding the 2216th gene could correctly identify about 20% of the observations that cannot be correctly classified by using only the 1184th gene. In fact, with these two genes, we reach a 100% overall CCP value and obtain a perfect classification for the three categories. The empirical distribution of all NRI values is shown in the left panel of Figure 1. The mode of the distribution is around 0.05, indicating a 5% improvement over the existing marker.

We have used the IDI for this data set and obtained the same results. At the first step, the 1184th gene was selected with the highest IDI 0.6364, while at the second step the 2216th genes was added for providing the most IDI increase 0.3906 (SE 0.0696). No further improvement in the IDI can be attained

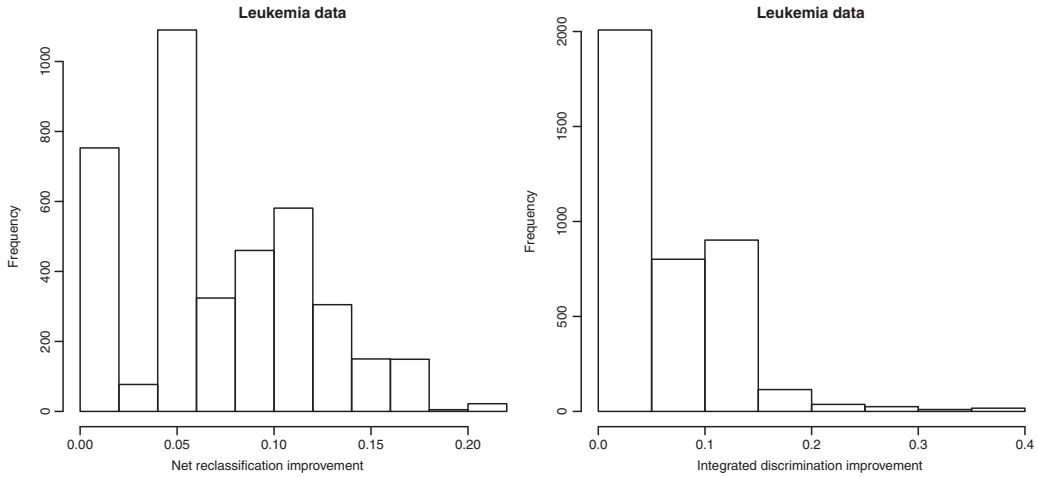


Fig. 1. Empirical distribution of NRI and IDI values for all gene expressions of the Leukemia example.

with additional genes. The empirical distribution of all IDI values is shown in the right panel of Figure 1. The distribution for the IDI is more skewed than that for the NRI.

Many authors examined this data set using various classification methods such as machine learning (Golub *and others*, 1999; Furey *and others*, 2000; Guyon *and others*, 2002), threshold circuits (Albrecht *and others*, 2003), rigid regression (Li and Yang, 2005), and stochastic search (Albrecht, 2007). The numbers of gene expression levels used in their studies were all greater than 2. The most similar previous results may be found in Li and Yang (2005) and Albrecht (2007), where only three genes were needed to achieve the same accuracy. Our findings based on NRI and IDI selection appear to be a further improvement from the existing analyses.

We may further use this example to compare across different methods. Besides the NRI and the IDI, a common ROC-based measure for multicategory classification is the HUM (Li and Fine, 2008). The results of using the HUM are the same as those obtained as using the NRI and the IDI: selecting the 1184th gene with the highest HUM value 0.8116 at the first step, and selecting the 2216th gene with the highest HUM improvement at the second step. This results in the same two genes being selected (in the same order) to give a 100% HUM in the final model. However, in general the agreement between the three approaches varies because they focus on different aspects of the model accuracy. The sample correlation coefficient between the NRI and the IDI values at the second selection step was 0.747. The scatter plot of the IDI versus the NRI is given in the upper left panel of Figure 2, displaying an increasing association. The correlation between HUM and NRI was only 0.627 and the scatter plot of HUM versus NRI is given in the upper right panel of Figure 2. The correlation between HUM and IDI was 0.859, showing a strong dependence pattern in the lower panel of Figure 2. All correlations are positive and strong, exceeding 0.60, with the association between IDI and HUM suggesting that, at least in this example these measures provide the most similar assessment of improvements in leukemia classification accuracy.

5. DISCUSSION

While in the numerical studies, multinomial logistic regression was employed for constructing the probability assessment, in practice, any procedure providing such assessments could be used. The simplicity of the logistic analysis is appealing and is theoretically supported by the recent work of Delaigle and Hall

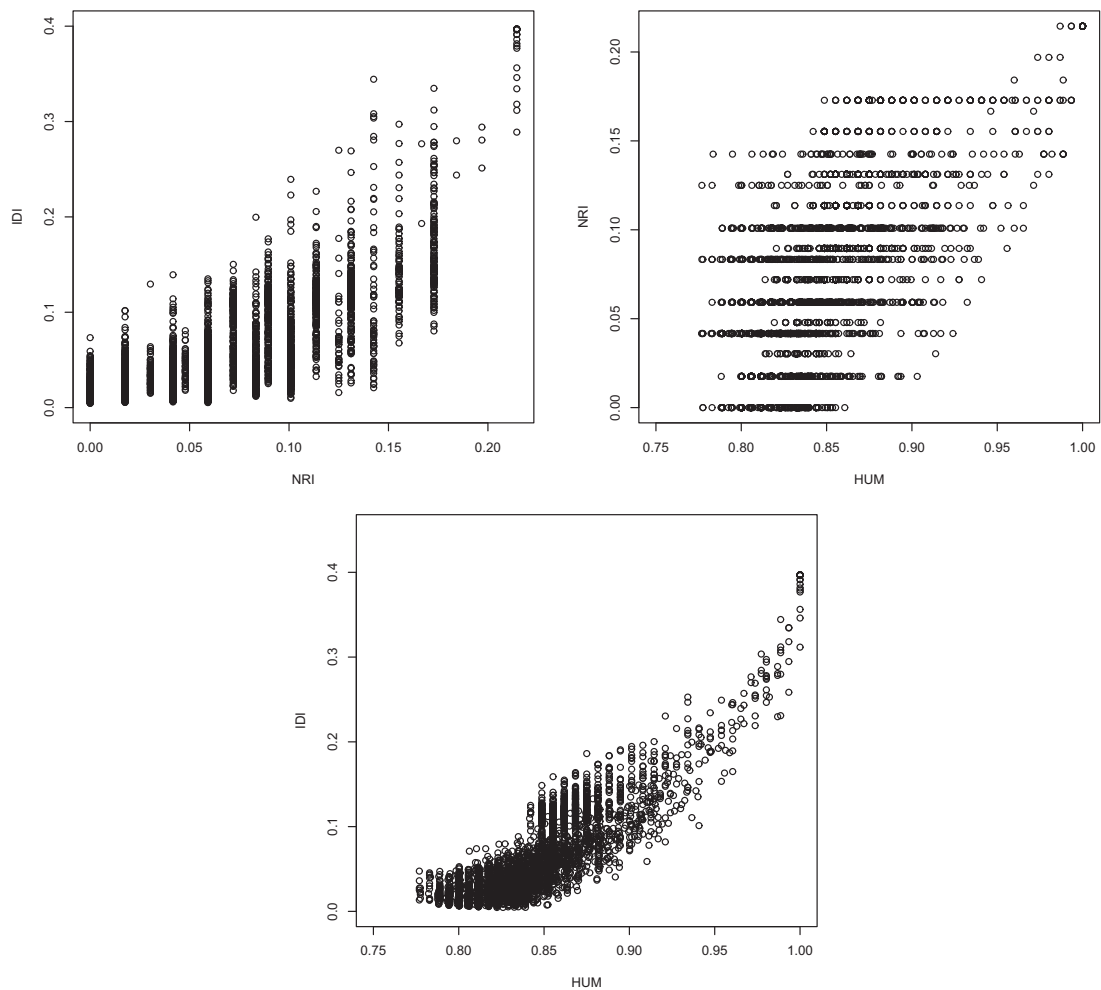


Fig. 2. Scatterplots of IDI versus NRI, NRI versus HUM and IDI versus HUM for gene expressions in the leukemia example.

(2012), who established that optimal classification can be achieved with a linear method for Gaussian data. Considering that normal distribution is perhaps the most prevalent for discrimination and classification (Pencina and others, 2012), the linear combination approach, thus, may provide satisfactory performance in a wide range of real data applications in the biomedical sciences.

Besides the multinomial logistic regression model, we have also experimented with the support vector machine as an alternative classifier and found its performance to be comparable with that of multinomial logistic regression in simulations. Code implementing such analyses is available at the first author's website mentioned earlier. An advantage of logistic regression beyond classification is that the coefficients may be easily interpreted and yield insight into the markers' effects on the response. Another observation is that the sampling behavior of NRI and IDI estimates seems less stable for the support vector machine when the sample size is small.

It is important to note that the evaluation of the NRI and the IDI must be based on correctly calibrated models, especially when the old and new models are not nested. In numerical studies, we have observed that the probability assessments from incorrectly calibrated non-nested models could be surprisingly different and that calculations based on such quantities may not yield reasonable results. That is, spurious improvements in the NRI and the IDI may be achieved when either one or both of the models fit poorly. In such scenarios, the improvements are confounded by the lack of fit of the models, potentially leading to incorrect conclusions about the predictive values of added biomarkers. In general, careful consideration of model calibration, including goodness-of-fit diagnostics, is needed to ensure adequate model fit prior to model comparisons based on the NRI and the IDI.

SUPPLEMENTARY MATERIAL

Supplementary material is available at <http://biostatistics.oxfordjournals.org>.

ACKNOWLEDGMENTS

The authors are grateful to an Associate Editor and Referee for helpful comments and suggestions. *Conflict of Interest*: None declared.

FUNDING

This work was supported by grants from Academic Research Funding R-155-000-130-112 and National Medical Research Council NMRC/CBRG/0014/2012.

REFERENCES

- ALBRECHT, A. (2007). Stochastic local search for the feature set problem, with application to microarray data. *Applied Mathematics and Computation* **183**, 1148–1164.
- ALBRECHT, A., VINTERBO, S. A. AND OHNO-MACHADO, L. (2003). An epicurean learning approach to gene-expression data classification. *Artificial Intelligence in Medicine* **97**, 245–271.
- COX, D. R. AND WERMUTH, N. (1992). A comment on the coefficient of determination for binary response. *The American Statistician* **46**, 1–4.
- DELAIGLE, A. AND HALL, P. (2012). Achieving near-perfect classification for functional data. *Journal of the Royal Statistical Society: Series B* **74**, 267–286.
- FUREY, T. S., CRISTIANINI, N., DUFFY, N., BEDNARSKI, D. W., SCHUMMER, M. AND HAUSSLER, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16**, 906–914.
- GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLIER, H., LOH, M. L., DOWNING, J. R., CALIGIURI, M. A. and others. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.
- GUYON, I., WESTON, J., BARNHILL, S. AND VAPNIK, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning* **46**, 389–422.
- KERR, K. F., MCCLELLAND, R. L., BROWN, E. R. AND LUMLEY, T. (2011). Evaluating the incremental value of new biomarkers with integrated discrimination improvement. *American Journal of Epidemiology* **174**, 364–374.

- KRENN, V., MORAWIETZ, L., BURMESTER, G. R., KINNE, R. W., MUELLER-LADNER, U., MULLER, B. AND HAUPL, T. (2006). Synovitis score: discrimination between chronic low-grade and high-grade synovitis. *Histopathology* **49**, 358–364.
- LI, J. AND FINE, J. P. (2008). ROC analysis for multiple classes and multiple categories and its application in microarray study. *Biostatistics* **9**, 566–576.
- LI, J. AND FINE, J. P. (2010). Weighted area under the receiver operating characteristic curve and its application to gene selection. *Applied Statistics* **59**, 673–692.
- LI, F. AND YANG, Y. (2005). Analysis of recursive gene selection approaches from microarray data. *Bioinformatics* **21**, 3741–3747.
- MA, S. AND SONG, X. (2011). Ranking prognosis markers in cancer genomic studies. *Briefings in Bioinformatics* **12**, 33–40.
- MENARD, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statisticians* **54**, 17–24.
- MOSSMAN, D. (1999). Three-way ROCs. *Medical Decision Making* **19**, 78–89.
- NAKAS, C. T. AND YIANNOUTSOS, C. T. (2004). Ordered multiple-class ROC analysis with continuous measurements. *Statistics in Medicine* **23**, 3437–3449.
- PENCINA, M. J., D'AGOSTINO SR, R. B., D'AGOSTINO JR, R. B. AND VASAN, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine* **27**, 157–172.
- PENCINA, M. J., D'AGOSTINO SR, R. B. AND DEMLER, O. V. (2012). Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvements for normal variables and nested models. *Statistics in Medicine* **31**, 101–113.
- PENCINA, M. J., D'AGOSTINO SR, R. B. AND STEYERBERG, E. W. (2011). Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in Medicine* **30**, 11–21.
- PEPE, M. S., FENG, Z. AND GU, J. W. (2008). Comments on 'Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond' by M. J. Pencina and others. *Statistics in Medicine* **27**, 173–181.
- PEPE, M. S., JANES, H., LONGTON, G., LEISENRING, W. AND NEWCOMB, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology* **159**, 882–890.
- SHIU, S. Y. AND GATSONIS, C. (2012). On ROC analysis with non-binary reference standard. *Biometrical Journal* **54**, 457–480.
- SLANSKY, E., LI, J., HAUPL, T., MORAWIETZ, L., KRENN, V. AND PESSLER, F. (2010). Quantitative determination of the diagnostic accuracy of the synovitis score and its components. *Histopathology* **57**, 436–443.
- TJUR, T. (2009). Coefficients of determination in logistic regression models—a new proposal: the coefficient of discrimination. *The American Statistician* **64**, 366–372.
- ZHOU, X. H., OBUCHOWSKI, N. A. AND MCCLISH, D. K. (2002). *Statistical Methods in Diagnostic Medicine*. New York: Wiley.

[Received May 25, 2012; revised October 21, 2012; accepted for publication October 22, 2012]